Applied Data Science Capstone

The Battle of Neighborhoods

(Course 9 of IBM Data Science Professional Certificate Specialization)

By: Neeraj Tripathi

20 September 2020

Fit Hub Gym Center

Chicago, USA

Business Problem

To select a suitable location for opening up a Gym/Fitness Center in the city of Chicago, Illinois-USA.

Data

1. Data for neighborhoods of chicago along with their zip codes
https://www.seechicagorealestate.com/chicago-zip-codes-by-neighborhood.php

2. Data for all zip codes that belong to State of Illinois, USA
https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/download/?format=csv&refine.state=IL

3. Venue data for the above neighborhoods using Foursquare API

Methodology

1. Collect the data of all the zip codes belonging to Illinois.
2. Collect the data of all zip codes and corresponding neighborhoods of Chicago
3. Merge the datasets to obtain the neighborhood data of Chicago along with coordinates.

4. Visualize the neighborhoods on map using folium library to check the correctness of coordinates.

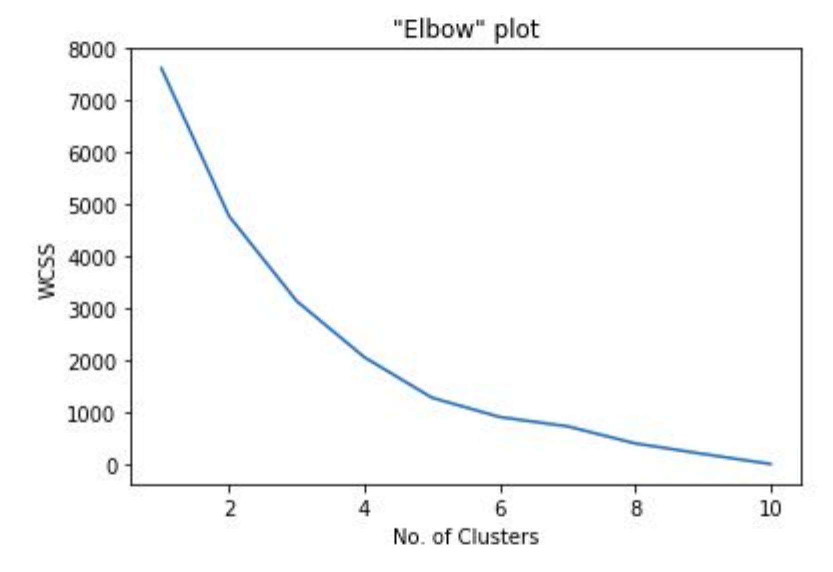
5. Obtain the venue data for each neighborhood present in the dataset using Foursquare API.

6. Preprocess the dataset, group similar venue categories into a single broad category. Drop the non required columns. Exploratory data analysis.

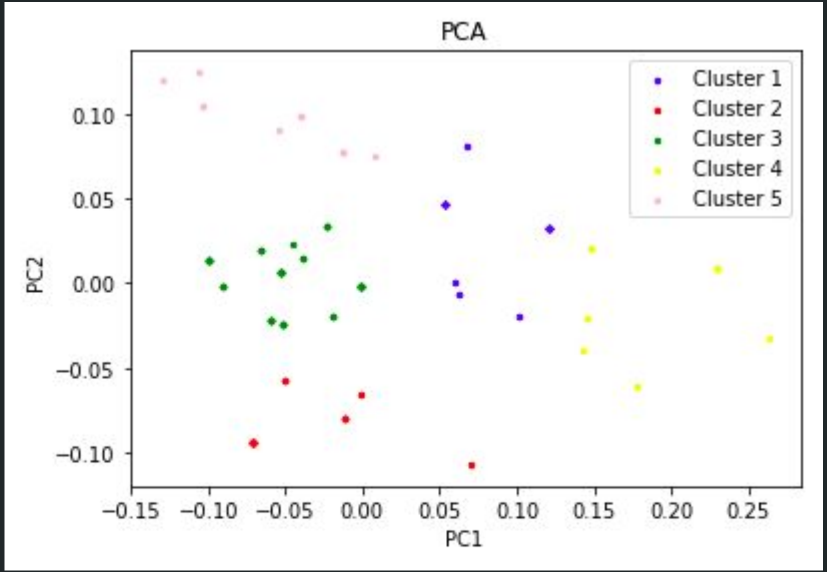
7. Find optimal number of clusters using “elbow method” and KMeans clustering algorithm.

8. Reduce the dimensions of the neighborhood venue frequency dataset to 2 using Principal Component Analysis (PCA).

9. Apply KMeans clustering algorithm on this reduced dataset and visualize the clusters to look for any inconsistencies.



1. Elbow plot to find optimal number of clusters



2. Plot reduced features (after PCA and clustering)

10. Inspect the number of gyms in the neighborhoods clusterwise by plotting all the gyms belonging to any cluster on the map.

11. Display top few neighborhoods along with the number of gyms for each cluster label.

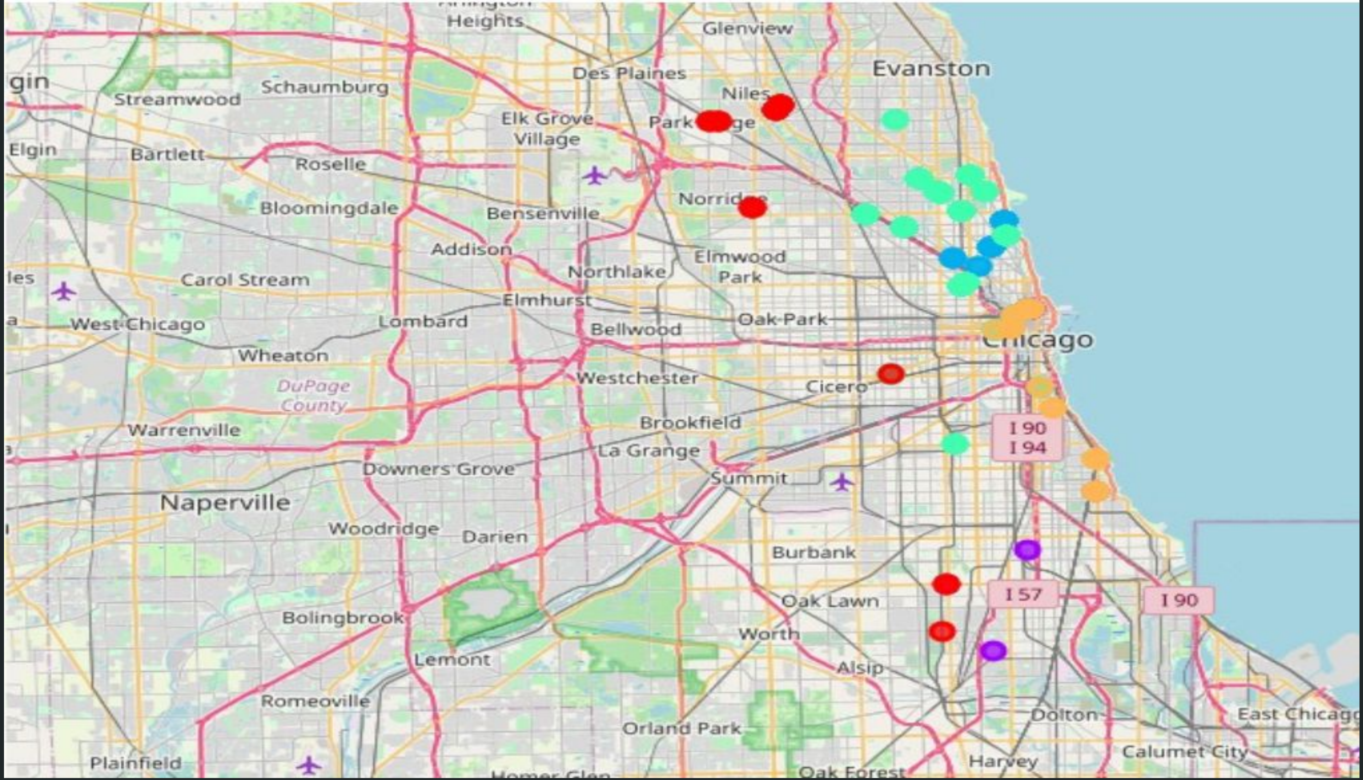
Results

1. The neighborhoods of Chicago can be grouped into 5 clusters based on the 10 broad categories described in this project.
2. Average number of gyms/fitness centers in the neighborhoods that belong to each cluster is different.

Cluster Label	**Average number of Gyms in a neighborhood**
0	2 - 4
1	5 - 6
2	3 - 5
3	2 - 4
4	1 - 3

Gyms plotted on map with colors indicating the “cluster labels”.

Blue, purple, green, orange and red colors represent gyms belonging to Cluster 0, 1 2, 3 and 4 respectively.



Observations

1. On the map, we can see that more and more gyms are located in the northern parts of the city.
2. Cluster 1 and Cluster 2 (which are the clusters where more than 50% of neighborhoods having more than 6 gyms and 3 gyms respectively) also fall mostly towards the northern parts, indicating a higer level of awareness among people towards health as well as higher competition in gym business in these areas.

Conclusion

1. Aggressive strategy: in Cluster 1 or Cluster 2 neighborhoods

2. Medium risk strategy: in Cluster 0 or Cluster 3 neighborhoods

3. Low risk strategy: in Cluster 4 neighborhoods